DUPLICATE ALSO

Numerical **Weather Prediction**

Forecasting Research
Scientific Paper No. 60

Multi-Model Multi-Analysis Ensembles in Quasi-Operational Medium Range Forecasting

by

K.R. Mylne, R.E. Evans and R.T. Clark

November 2000

ORGS UKMO F

National Meteorological Library
FitzRoy Road, Exeter, Devon. EX1 3PB

The Met.Office

Excelling *in weather services*

Forecasting Research
Scientific Paper No. 60

Multi-Model Multi-Analysis Ensembles in Quasi-Operational Medium Range Forecasting

by

Kenneth R. Mylne, Ruth E. Evans and Robin T. Clark*
The Met. Office,UK

(* *now at Meteo-France*)

November 2000

(Submitted to the *Quarterly Journal of the Royal Meteorological Society*)

The Met. Office
NWP Division
Room 344
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom

© Crown Copyright 2000

Permission to quote from this paper should be obtained from the above Met. Office division

Please notify us if you change your address or no longer wish to receive these publications.

Tel: 44 (0)1344 856245 Fax: 44(0)1344 854026 e-mail: jsarmstrong@meto.gov.uk

Multi-Model Multi-Analysis Ensembles in Quasi-Operational Medium-Range Forecasting

Kenneth R. Mylne, Ruth E. Evans and Robin T. Clark*

The Met Office, UK

*(*now at Meteo-France, France)*

(Submitted to the *Quarterly Journal of the Royal Meteorological Society*, November 2000)

SUMMARY

Ensemble prediction systems (EPS) for medium-range forecasting attempt to account for uncertainty in NWP forecasts by sampling the probability density function of future atmospheric states. Forecast uncertainty derives from uncertainty in both the analysed initial conditions (analysis errors) and in the forecast evolution (model errors). Current operational systems are primarily based on sampling the analysis errors through initial condition perturbations, with at best only limited sampling of model errors. One approach to sampling model errors, and also to widening the sampling of analysis errors, is to include more than one NWP model, and more than one operational analysis to which perturbations are added, in the ensemble system. Previous work has demonstrated from a small number of case studies that this multi-model multi-analysis ensemble (MMAE) approach can perform significantly better than a single-model system such as the EPS run by the ECMWF (European Centre for Medium-Range Weather Forecasts). In this study a MMAE ensemble was created by combining the ECMWF ensemble with an ensemble using the Met Office model and analysis, and was run daily for a year to assess the benefits over a larger, quasi-operational sample of forecasts. The results are compared with the operational ECMWF EPS which includes the latest upgrades, including stochastic physics which makes some allowance for uncertainty due to model errors. Results show that for both probabilistic forecasts (assessed by Brier Skill Scores and Relative Operating Characteristics) and deterministic forecasts based on the ensemble mean (assessed by Root-Mean Square Errors) the MMAE has increased forecast skill relative to the EPS. These improvements are obtained with no overall increase in ensemble size. Ensemble spread is also greater in the MMAE, and the increased skill is believed to be due to the additional model producing solutions which are more synoptically different than those produced by a single model ensemble. Benefits of the MMAE vary both in time and with geographical region, depending on which individual ensemble system performs better in particular synoptic situations. It is found that the MMAE almost always performs as well as the better individual ensemble, and on occasions better than either.

KEYWORDS: Ensemble Prediction; Multi-model Ensembles; Medium-Range Forecasting.

1. INTRODUCTION

Ensemble prediction systems for medium-range forecasting are designed to attempt to account for uncertainty in forecasts by sampling the pdf (probability density function) of possible future atmospheric states. The non-linear nature of the evolution both of the atmosphere and of the numerical models which attempt to simulate it, means that small errors either in the analysis of the initial conditions or in the formulation of the model can be rapidly amplified to give large differences between forecast and reality. In order to take account of all the uncertainties in estimating the forecast pdf, an ensemble system must attempt to sample both the analysis errors and the model formulation errors. Historically, early operational ensemble forecasting systems were designed to account only for analysis errors, by adding perturbations to the control analysis to generate ensemble members. Molteni *et al* (1996) and Toth and Kalnay (1993) describe the

original ensemble systems at the ECMWF (European Centre for Medium-Range Weather Forecasts) and NCEP (US National Centers for Environmental Prediction) respectively. No explicit account was taken of model errors, although their importance was recognised: Molteni *et al* (1996) attributed some limitations in the performance of the ECMWF ensemble, particularly with respect to transitions between different flow regimes, to systematic errors in the model climatology.

Ensemble systems using only initial condition perturbations typically display insufficient spread in the ensemble, characterised by an excessive proportion of verifying observations falling outside the range of forecast values. In an ideal ensemble the ensemble spread should, on average, equal the error of the ensemble mean (EM). In the case of the ECMWF ensemble the initial condition perturbations are scaled to achieve this after 48 hours of the forecast (T+48), but spread is too small later in the forecast period (Richardson, 2000). In practical terms for forecasting this results in many events not being predicted by any members of the ensemble, and in probability forecasts not giving reliable probabilities. Ensemble research is now increasingly looking at methods of taking account of other uncertainties in the forecast process. For example, Buizza *et al* (1999) have incorporated some elements of model uncertainty into the operational ECMWF ensemble by adding stochastic perturbations to model physics within individual ensemble members; Houtekamer *et al* (1996) run a number of separate versions of their model in an ensemble, with different parametrization schemes.

Despite the inclusion of the stochastic physics in the ECMWF ensemble, it is a common observation of forecasters in the Met Office that EPS members follow too closely to the solution of the ECMWF high-resolution deterministic model (or of the EPS control) - in other words that the EPS does not spread sufficiently to incorporate the full range of uncertainty. They observe subjectively that model forecasts from other NWP centres are synoptically more different from the ECMWF model than are the EPS members.

Another approach to improving ensemble performance, making use of the different solutions of other NWP systems observed by the forecasters, is to create a multi-model multi-analysis ensemble by combining ensembles run using different models. As well as the additional models, the use of more than one analysis (to which perturbations are added) may provide further unstable directions for ensemble spread not present with a single analysis. The use of multi-model multi-analysis ensembles is now becoming standard practice in seasonal-range forecasting where the benefits were demonstrated in the PROVOST project (Graham *et al* 2000) and will be used extensively in the forthcoming European DEMETER project. Multi-model multi-analysis techniques are also being used increasingly for short-range ensemble forecasting. Stensrud *et al* (1999) found that the inclusion of two different models in an ensemble for short-range forecasting "assists in increasing the ensemble spread significantly". Mullen *et al* (1999) found that the combined effect of uncertainties in model physics and in the initial state provided a suitable means for creating a short-range ensemble for QPF (quantitative precipitation forecasting), with the inclusion of different forms of cumulus cloud parametrisation schemes within a mesoscale model ensemble. In medium-range forecasting, Evans *et al* (2000) have conducted a number of case studies by combining ensembles from the ECMWF and Met Office Unified Model (UM) NWP systems, using each model with the analyses from its home centre. They concluded that the joint ensemble significantly outperforms either individual ensemble, giving a gain in predictability of the order of one day over the ECMWF Ensemble Prediction System (EPS), with better coverage of the observations and evidence that the UM ensemble can include synoptically valuable information not included in the EPS. One limitation of this work was that it only considered a number of case studies, selected in part because of poor EPS performance. Also the model versions used were of lower resolution than the current operational

EPS. In order to overcome these limitations, a multi-model multi-analysis ensemble (MMAE) was implemented at ECMWF in a quasi-operational mode, using the current operational versions of both ECMWF and Met Office models. This system was run daily for a year from October 1998 to October 1999. This paper reports the results from these experiments. A number of diagnostic verification tools were employed in the analysis and it is not possible to describe all results in detail. Detailed results will be discussed using the Brier Skill Score, which gives a good overall assessment of the probabilistic skill of an ensemble system, and results using other diagnostics will be summarised more briefly. Some results from this work have also been discussed by Richardson (2000) who compares the benefits of the MMAE system with those from multi-analysis ensembles using just a single model.

2. THE MULTI-MODEL MULTI-ANALYSIS ENSEMBLE (MMAE) SYSTEM

The multi-model multi-analysis ensemble was created by combining the 51-member ECMWF operational EPS (Ensemble Prediction System) with a 27-member ensemble created using the Met Office's Unified Model (UM) NWP system.

The original implementation of the ECMWF EPS was described by Molteni *et al* (1996); the system was subsequently upgraded in 1996 (Buizza *et al*, 1998) when the ensemble was increased from 33 to 51 members and the model resolution was increased from T63L19 to T_L159L31. (T63L19 is horizontal spectral triangular truncation T63 with 19 vertical levels; in T_L159L31 the subscript L indicates a 'linear grid' option, with T_L159 equivalent to a triangular truncation of T106; Buizza *et al* 2000.) The current version of the EPS used in these experiments thus consists of a control integration of the ECMWF model run at T_L159L31 resolution from the unperturbed ECMWF analysis, plus 50 integrations run from perturbed analyses. Perturbations are generated from the so-called singular vectors (SVs) of the tangent forward model (Buizza and Palmer, 1995) which produce the fastest energy growth in the early stages of the forecast. Each of 25 perturbations is added to, and subtracted from, the analysis, to produce a pair of perturbed analyses which are used as initial conditions for a pair of ensemble members. A number of upgrades to the operational EPS are described by Buizza *et al* (2000). It is important to note that these upgrades were installed prior to the start of the experiments reported in this paper. The last and most relevant upgrade was the inclusion of a simulation of random model errors, by the application of stochastic perturbations to the tendency due to parametrized physical processes (Buizza *et al*, 1999). This upgrade, referred to hereafter as "stochastic physics", was installed operationally on 21 October 1998; results presented in this paper are all taken from after this date. The EPS is run once per day from the 12UTC analysis out to 10 days (240h) ahead.

To implement the UM ensemble, the UM (Cullen, 1993) was installed on the computer system at ECMWF. The version used was the same as the operational global NWP model run at the Met Office at that time, except that the horizontal resolution was reduced to 288x217 grid-points, or 1.25° E-W by 0.833° N-S, with 30 vertical levels. This gave a similar resolution to the operational EPS model at T_L159L31. It is important to note that the UM is fundamentally different from the ECMWF model in that both the dynamical core of the model and most of the physics parametrizations are different and independent of each other. This means that while both NWP systems are amongst the most skilful available, they are likely to have different strengths and weaknesses. Initial conditions for the UM ensemble were created by adding the EPS perturbations to the Met Office's operational analysis. Due to limited computing resources, the UM ensemble was run with 27 members (control plus the first 13 pairs of perturbed analyses), compared to the EPS which has 51 members (control plus 25 pairs). Like the EPS, the UM

ensemble was run daily up to 10 days ahead using the 12UTC Met Office analysis. The use of the EPS perturbations is likely to produce a less optimal performance from the UM ensemble than from the EPS, since the perturbations are specifically calculated to generate maximum ensemble growth using the ECMWF model. However resources were not available to develop a method of calculating perturbations optimised for the UM. Since the perturbations are primarily determined by the current state of the atmosphere, it is believed that, although less than optimal, the EPS perturbations should still provide effective ensemble growth with an alternative model such as the UM.

In the analysis of results, output from the UM ensemble is combined with EPS output to create the MMAE ensembles. The MMAE therefore uses two completely independent models and two independent operational analyses, but only a single set of analysis perturbations. Two versions of MMAE are used: (i) a full version combining the 27 UM members with all 51 members of EPS to give a 78 member joint ensemble (referred to hereafter as EEUU78) and (ii) a reduced version consisting of the 27 UM members combined with the corresponding 27 members of the EPS (EEUU54). The latter provides a useful comparison with the EPS for assessing the benefits of the MMAE system, as the total number of members is similar. Any benefits achieved are therefore due to the use of the additional model and analysis, and are not due simply to increasing the number of ensemble members. Since the ensemble size is similar, any benefits can be achieved without any significant increase in computing costs.

3. ANALYSIS OF EXPERIMENTAL RESULTS

Analysis of results from the MMAE experiments involves processing very large quantities of data, and it has therefore not been possible to analyse the entire year in which the MMAE was run. Verification results are presented here for a 3-month period in winter from December 1998 to February 1999 inclusive (hereafter referred to as DJF), and a 3-month period in early summer from May to July 1999 (MJJ). (It is not possible to use the more conventional JJA period for the summer, as processing was not completed for August 1999.) A total of 75 days were analysed in DJF and 85 days in MJJ, making a much larger dataset than has been analysed in the previous case-study work. It was also only possible to calculate a relatively small set of diagnostics. In particular probabilistic skill scores were only calculated for forecasts of whether parameters are above or below the climatological normal. Limited resources prevented analysis of forecast skill for more extreme events.

Four meteorological fields from the ensembles were verified: mean-sea-level pressure (MSLP), 500hPa geopotential height (H500), 850hPa temperature (T850) and 24-hour precipitation accumulations (P24). For verification all fields were interpolated onto a common grid. Verification was carried out at UM ensemble resolution of 1.25x0.83 degrees for MSLP, H500 and P24, and at approximately quarter resolution (5.0x3.0 degrees) for T850. Fields were verified every 12 hours through the forecast range from T+12 to T+240 for MSLP and H500, and every 24 hours from T+24 to T+240 for P24 and T850. Results are mostly presented here for MSLP since this field is particularly relevant to medium-range forecasters using the ensembles, with results for other parameters described for comparison. Scores were calculated for several geographical regions as given in table 1. Results are presented primarily for the North Atlantic and Europe region, with other regions used for comparison where appropriate.

Verification was performed using the ECMWF analysis as the verifying "truth" (or in the case of precipitation, the 0-24h short range forecast accumulation from the ECMWF high-resolution model was used). Some more limited verification was also carried out using the Met Office analysis and a consensus analysis, the latter being the mean of the ECMWF and Met Office

analyses. The effect of the choice of verifying analysis will be discussed in section 4.2, but overall results are mostly independent of which analysis was used beyond about 48h into the forecast range. Since the EPS is designed for medium-range forecasting beyond 48h, differences in the short-range are not important. Except where explicitly discussing the effect of the verifying analysis, results presented in this paper will all be verified against the ECMWF analysis.

Region	**Description**
Europe	30-75N, 20W-45E
North Atlantic and Europe	30-75N, 65W-45E
Northern Hemisphere Extratropics	30-80N
Southern Hemisphere Extratropics	30-80S

Table 1: Geographical regions used for analysis of results

Following normal convention (eg Wilks, 1995), comparison of the skill of different forecasts is normally done in terms of a skill score. For any verification diagnostic, X, the skill of a forecast relative to some reference forecast is given by

$$S = \frac{X_r - X_f}{X_r - X_p} \tag{1}$$

where X_f is the value of X for the forecast, X_r for the reference forecast and X_p for a perfect deterministic forecast. A skill score has a maximum value of 1 (or 100%) for a perfect forecast (X_f=X_p) and a value of zero for performance equal to that of the reference (X_f=X_r). S has no lower limit, with negative values representing poorer skill than the reference.

Normally the reference forecast used is a standard baseline such as persistence. For probability forecasts climatology is often used. In this case we are interested in the performance of the MMAE ensemble compared to the EPS, so the EPS forecast is used as the reference. Results in this paper will mostly be presented as skill scores using the EPS as reference.

4. ANALYSIS USING BRIER SKILL SCORES

The most commonly used verification diagnostic for probabilistic forecasts is the Brier Score, originally introduced by Brier (1950) and described in its modified standard form by Wilks (1995) as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2 \tag{2}$$

The Brier Score is essentially the mean square error for probability forecasts of an event, where f_i and o_i are forecast and observed probabilities respectively; o_i takes values of 1 when the event occurs and 0 when it does not. Brier Skill Scores (BSS) are calculated by using BS as X in eq. (1).

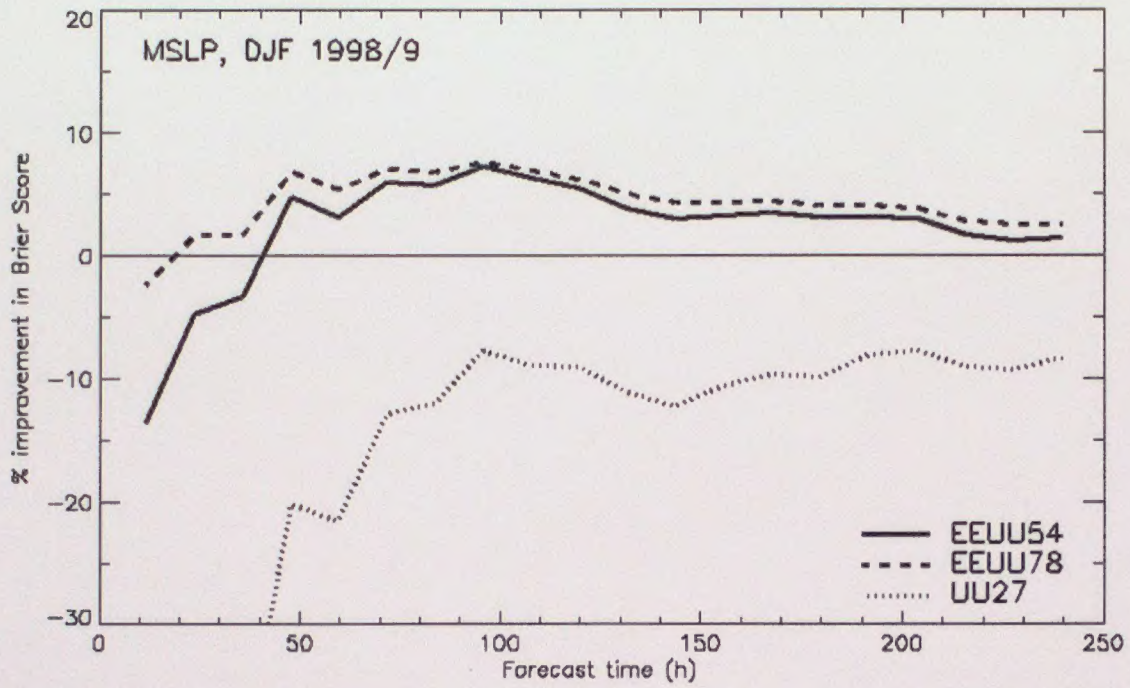
4.1 BRIER SKILL SCORES FOR NORTH ATLANTIC AND EUROPE

In Fig. 1 Brier Skill Scores of the two configurations of the MMAE are plotted against forecast lead-time for the North Atlantic and Europe in the (a) DJF and (b) MJJ periods. Also shown are the scores for the UM ensemble (UU27). Although the UM ensemble alone is considerably less skilful than the EPS, in winter (DJF) the EEUU54 gives a gain in skill of between 3 and 7% over the EPS for lead-times of T+48 to T+204. When Brier Scores are plotted (not shown), rather than skill scores, it can be seen that this improvement in skill from the MMAE equates to a gain in predictability of around 6 hours, and up to 12 hours beyond T+156. In summer (MJJ) EEUU54 gives 2-3% improvement over the EPS between T+72 and T+168. There is a noticeable diurnal fluctuation in the MJJ results, with greater MMAE benefit in forecasts verifying at 12UTC than in those at 00UTC. Over the first 2 days of the forecast, performance of the UM and MMAE ensembles appears poor in both seasons. This is believed to be due to using the ECMWF analysis for verification, and will be discussed in detail in section 4.2.

Similar results are obtained for geopotential height at 500hPa (H500). Figure 1(c) shows results for MJJ; results for DJF are almost identical to the MSLP scores in Fig. 1(a). The diurnal variation of the MMAE benefit in MJJ is even more pronounced at H500 than for MSLP, but it is in the opposite sense, with greater benefits for forecasts verifying at 00UTC. For P24 the gains from the MMAE are smaller than for MSLP and H500, but positive. Figure 1(d) shows results for DJF with up to 3% improvements in skill for EEUU54; in MJJ (not shown) the benefit of EEUU54 is only between 0 and 1%, although EEUU78 gives 1-3% gain for all lead-times. For T850 (not shown) the EEUU54 ensemble gives improvements of up to 3% beyond T+96 in DJF, but *BSS* are negative for shorter lead-times and for all lead-times in MJJ. However these results for T850 are found to be strongly influenced by the choice of verifying analysis. Reasons for this will be discussed in more detail in section 4.2, but results for T850 will not be considered further as they cannot be used to draw reliable conclusions.

There is little difference between results for the two MMAE ensembles, EEUU54 and EEUU78 in Fig. 1, although the larger ensemble does score consistently slightly higher. This shows that most of the benefits from an MMAE system are due to real gains in skill, and not simply to increasing ensemble size, which means that they can be achieved at little extra computing cost. There is slightly more difference for H500 in MJJ (Fig. 1c), when EEUU78 scored around 2% better than EEUU54 for all lead-times, and this was enough to keep the EEUU78 scores positive beyond T+120, whereas the EEUU54 score fell slightly negative for fields verifying at 12UTC. There is also slightly more difference for precipitation forecasts (eg Fig. 1d). This may be due to the much less smooth nature of precipitation fields which may mean that the quality of probability forecasts may be more sensitive to ensemble size than it is for smoother fields. For all fields the EEUU78 also gives more benefit in the short-range period, where it is noticeable that while the EEUU54 performs less well than the EPS, the skill scores for EEUU78 are positive (or close to zero) for all fields, even though the use of the ECMWF analysis causes the UM members to score poorly. The fact that EEUU78 scores are almost always positive suggests that adding additional members from another model/analysis system to the EPS does not significantly degrade the EPS performance, even where the independent performance of the additional members is less good, although clearly if large numbers of members were introduced from much less skilful NWP systems then the overall performance would be degraded.

(a)



(b)

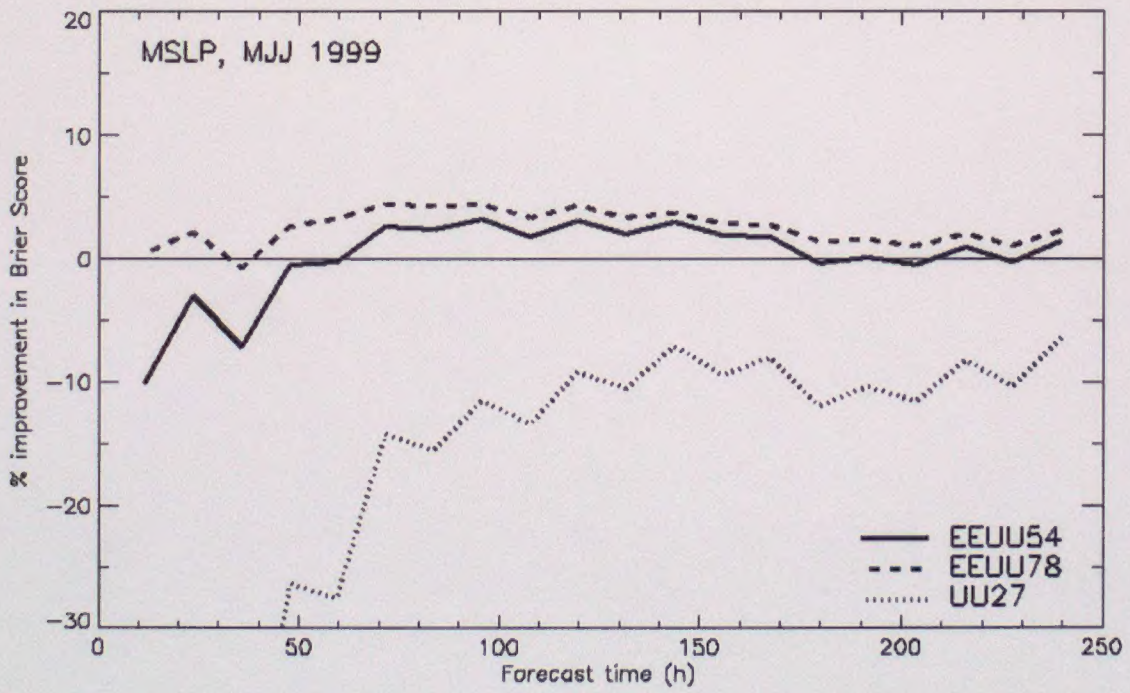
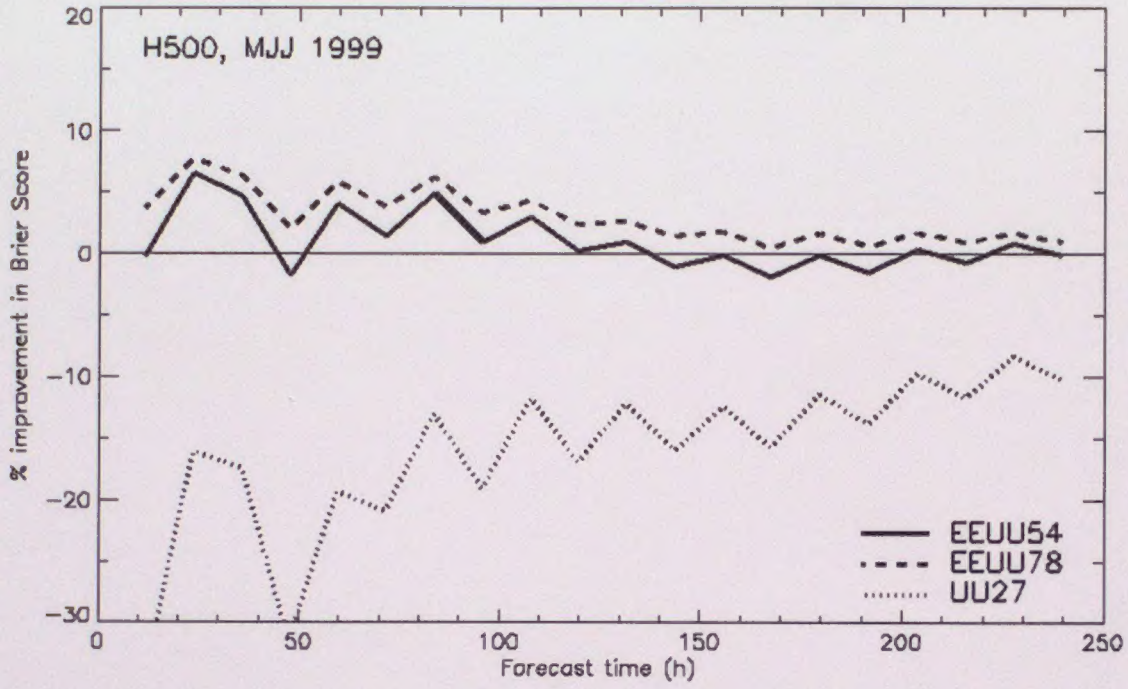


Figure 1: (contd overleaf)

(c)



(d)

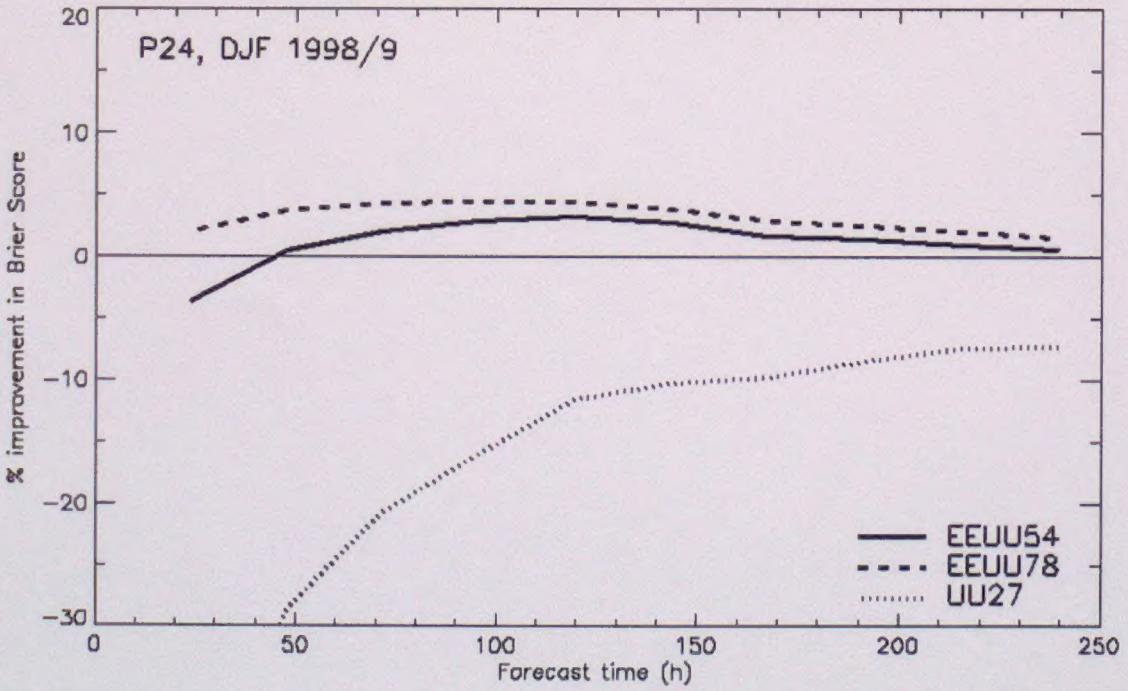


Figure 1: Brier Skill Scores relative to the EPS, calculated over the North Atlantic and Europe for (a) MSLP above normal in DJF (b) MSLP above normal in MJJ, (c) H500 above normal in MJJ and (d) P24 above normal in DJF. Different ensembles are as shown in the key: EEUU54 and EEUU78 are the two versions of MMAE ensembles as described in section 2; UU27 is the 27-member UM ensemble.

4.2 EFFECT OF THE CHOICE OF VERIFYING ANALYSIS

It appears from Fig. 1 that the UM ensemble is performing very poorly in the first 48 hours of the forecast, and this leads also to poor performance of the MMAE ensembles in this period, with negative skill scores. This result is in fact due to the use of the ECMWF analysis for verification. In Fig. 2(a) *BSS* are shown for EEUU54 from January 1999 using the Met Office and consensus analyses, as well as the ECMWF analysis, for verification. Using the Met Office analysis, the MMAE (and UM, not shown) ensembles verify better than the EPS in the first 48 hours; using the consensus analysis there is no poor performance of the EEUU54 in the first 48h. Operational analyses are generated in a cycle in which the background fields are provided by short-period forecasts, with the result that analysis fields are characterised by the short-period model biases of the model used. It is therefore to be expected that forecast fields produced by a different model, using its own analysis for initial conditions, will appear to verify less well in the early stages of the forecast. Later in the forecast period the forecast errors become dominated by synoptic evolution errors, and the effect of model biases in the analysis should become small, although differences in model biases may always give some benefit to the model whose analysis is used for verification. It can be seen from Fig. 2(a) that after 48 hours the difference due to the analysis reduces, and becomes relatively small by T+120. Fig. 2(b) shows similar graphs for the southern hemisphere. Here the effect of the choice of analysis is rather larger in the period T+48 to T+96. In the southern hemisphere the influence of the model used to create the analysis persists for longer than in the northern hemisphere because of the larger data-sparse areas. In these areas the analysis is dominated by the short-period forecasts produced by the model, as there are few observations to change the analysis from the model background fields.

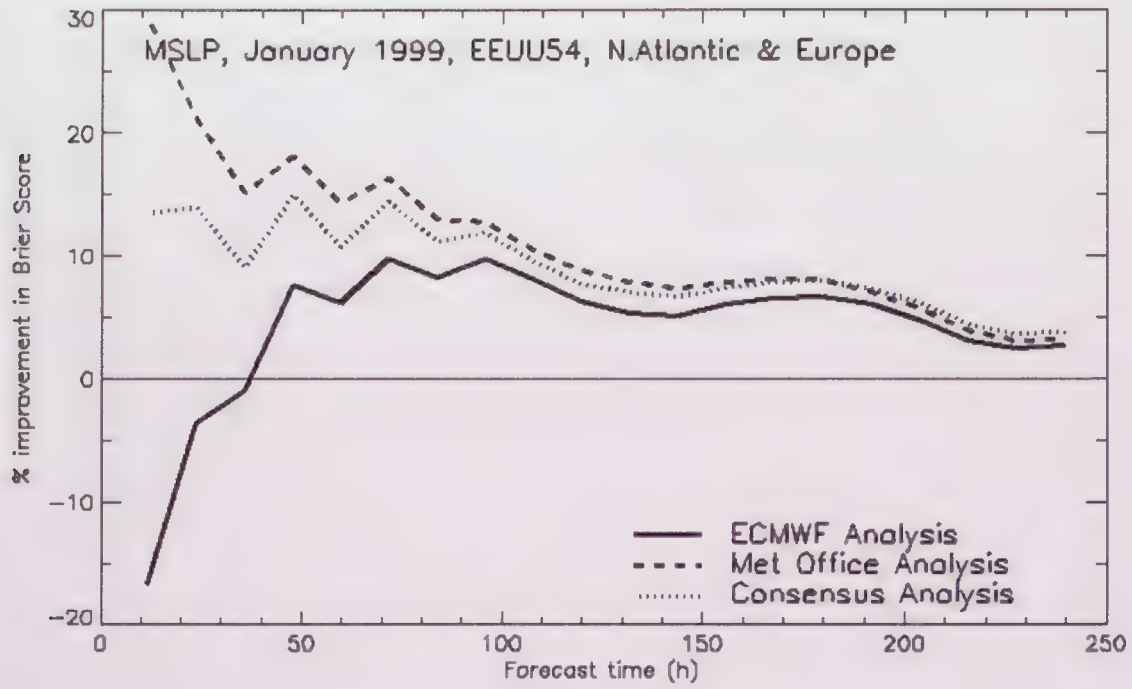
From Fig. 2 it can be seen that the apparent benefit from the MMAE is virtually independent of the choice of verifying analysis beyond T+120. Between T+48 and T+120 the apparent size of the benefit depends on the analysis used, and in some cases such as the southern hemisphere example shown in Fig. 2(b) this could affect the conclusions about whether the MMAE gives a positive benefit or not. Since the aim of the verification in the current experiments is to investigate benefits of the MMAE *relative to EPS*, the ECMWF analysis is used throughout the rest of this paper to avoid any risk of biasing results in favour of the UM or MMAE ensembles. Since the ensembles are designed for use on medium-range time-scales, beyond 48 hours, the short-range effects are not important to the conclusions.

One exception to these conclusions about the choice of verifying analysis is for T850 fields. It was noted above that in the summer the T850 forecasts from EEUU54 were poorer than EPS for all lead-times. If the Met Office analysis is used the EEUU54 verifies as better than the EPS at all lead-times. This is believed to be due to a known cold bias in the UM at 850hPa at all lead-times, which means that each model's forecasts will always verify better against its own analyses. Since the overall conclusions drawn would depend on the choice of verifying analysis in this case, T850 results will not be considered further.

4.3 TIME AND SPACE DEPENDENCE OF MMAE BENEFITS

Results have been given above for the two 3-month seasons DJF and MJJ. If the individual seasons are broken down into shorter periods there are considerable variations in the benefits of the MMAE. In Fig. 3 *BSS* results are shown for each month of the winter DJF season for MSLP from the EEUU54 ensemble over the North Atlantic and Europe. Clearly the benefit from the MMAE approach varies significantly from month to month during this period. Benefits were greatest in January with improvements in skill over the EPS of over 5% for all lead-times from

(a)



(b)

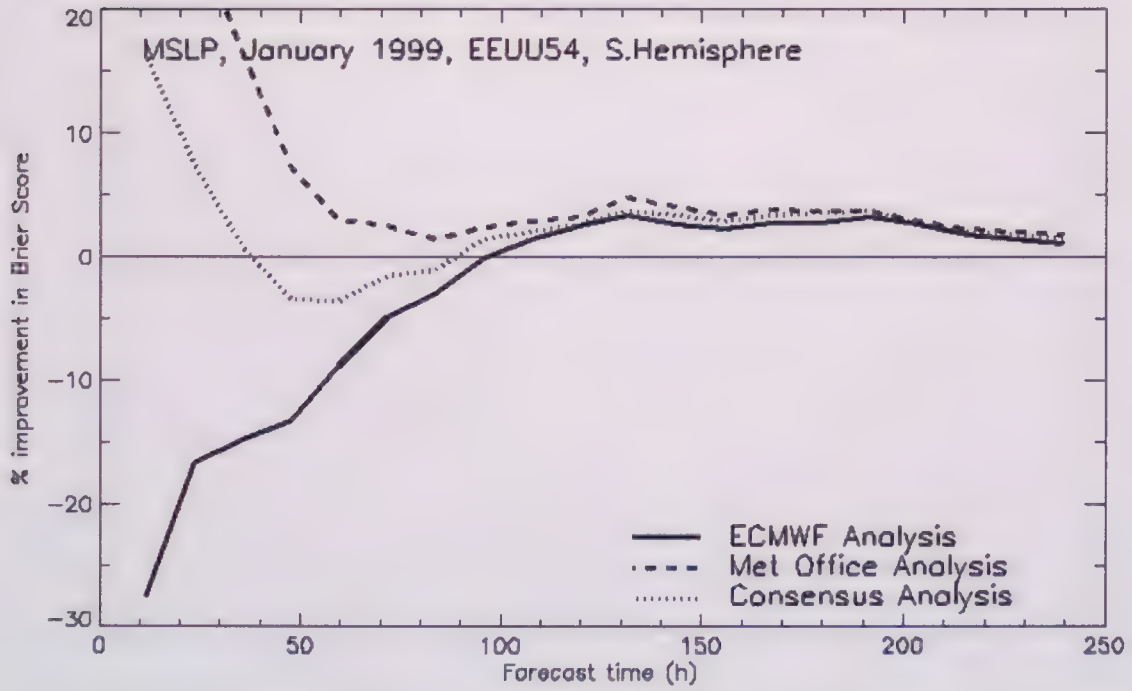


Figure 2: Brier Skill Scores relative to the EPS for MSLP from the EEUU54 ensemble, verifying against three different analyses as shown in the key, in (a) the N. Atlantic and Europe and (b) the Southern Hemisphere. Results shown are for January 1999; results for other months show very similar differences between the three analyses.

T+48 to T+204, and up to 10% around T+72 to T+96. In December and February the maximum gains were around 5%, and there was a small reduction in skill compared to the EPS beyond T+168 in December. (For EEUU78 (not shown) skill falls to zero beyond T+168, but is not significantly negative.) Examination of the scores for UU27 shows that the UM component of the MMAE had very similar skill to the EPS in January, whereas in December and February it was 5-15% poorer. This clearly contributed to the better skill overall of the MMAE in January. However it is important to note that while the EEUU54 ensemble performs better than the EPS when the UM ensemble performs well, it is not significantly worse than the EPS when the UM component performs poorly. There are some minor exceptions to this, such as in December beyond T+168, but these are far outweighed by the benefits gained at other times.

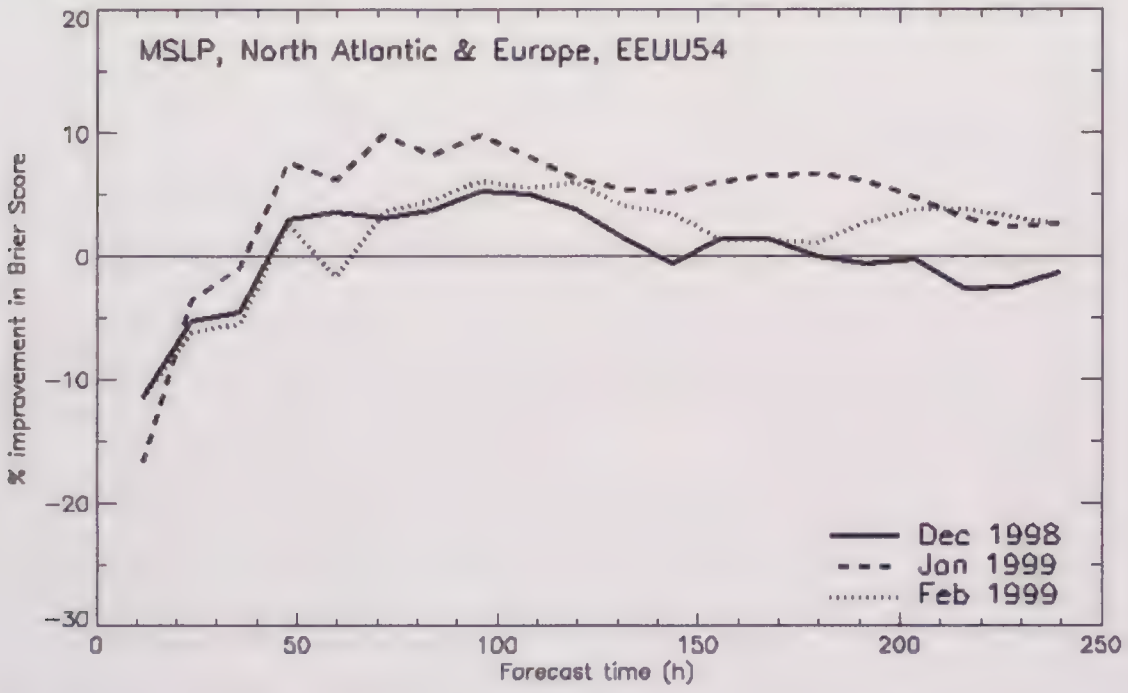


Figure 3: Brier Skill Scores relative to the EPS for MSLP from the EEUU54 ensemble, calculated over the North Atlantic and Europe for individual months in DJF.

The time dependence of the relative performances of the different ensembles can be broken down further. In Fig. 4 the Brier Scores for T+120 forecasts of MSLP for individual days are plotted for the EPS (EE) and the UU27 and EEUU54 ensembles, averaged over the North Atlantic and Europe area. It can be clearly seen that the relative performance of the individual ensembles varies from day to day. The points marked E indicate examples of days where the EPS performs better, and those marked U where the UM ensemble is better (noting that the Brier Score is negatively oriented with the lower score indicating better performance.) However on almost every day the EEUU54 joint ensemble performs as well as the better individual ensemble, and on occasion, such as the point at the extreme left end of the plot in Fig. 4, performs better than either individual one. Thus the inclusion of a second model into the ensemble frequently adds additional useful information, and importantly it does not significantly reduce the performance of an individual ensemble when it is performing well.

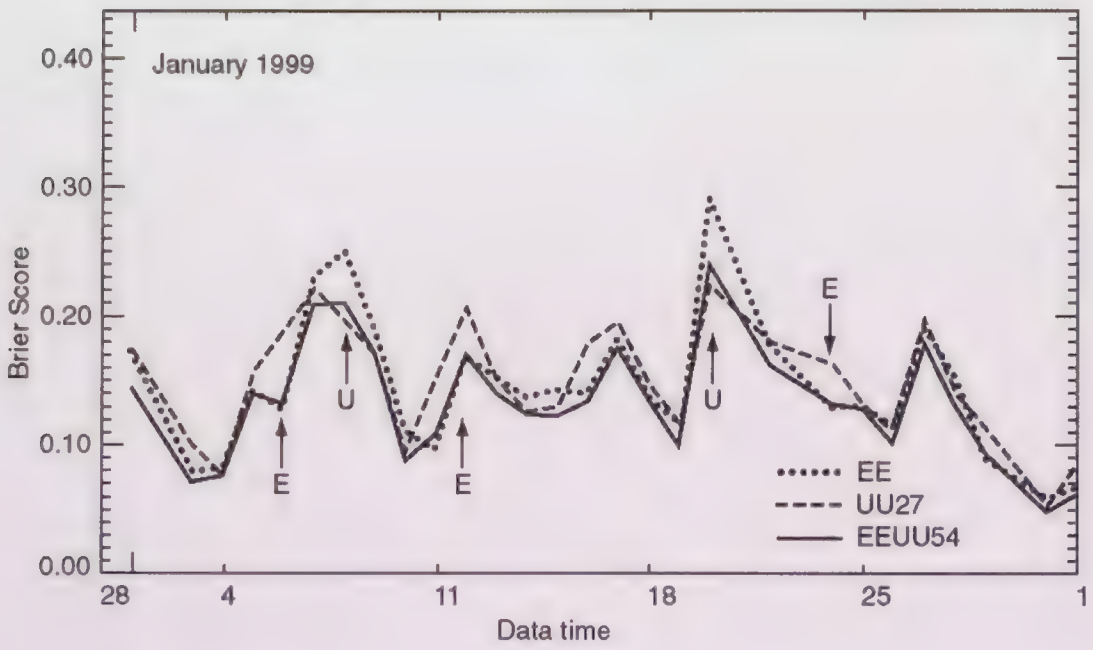
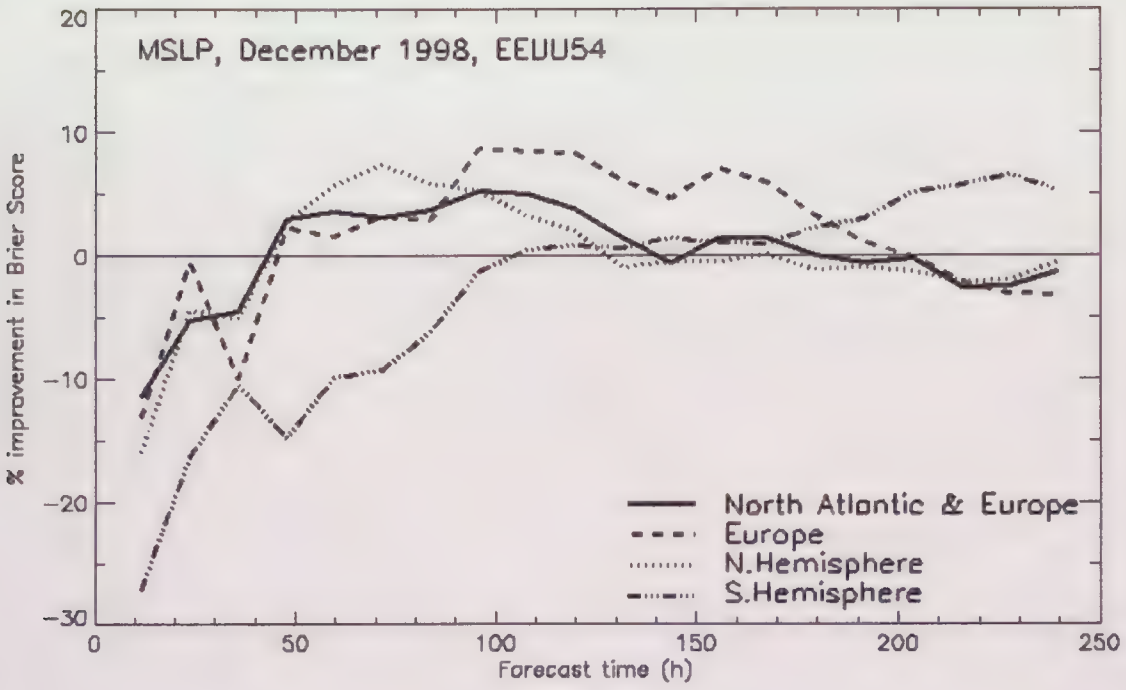


Figure 4: Brier Scores for individual days during January 1999, for 120-hour forecasts of MSLP above normal calculated over the North Atlantic and Europe. Different ensembles are as shown in the key, where EE is the EPS. Points marked E indicate examples of where the EPS is the better individual ensemble; points marked U where UU27 is better.

As well as looking at the variation in benefits from month to month, we can also look at the effect of varying the region over which scores are calculated. Figure 5 shows *BSS* for MSLP from EEUU54 over the four geographical regions used in (a) December 1998 and averaged over (b) the DJF and (c) MJJ seasons. From the December results it can be seen that although the benefits from the MMAE over the North Atlantic and Europe region were relatively small, as discussed above, over the smaller Europe region there was much more benefit, with *BSS* of 3-8% for all lead-times from T+48 to T+180. There were also larger gains in the period T+48 to T+96 over the whole of the Northern Hemisphere, which indicates that there must have been significant gains in other areas outside the North Atlantic and Europe region. These gains are all considerably larger than the negative skill scores which are observed for some regions at longer lead-times, indicating an overall benefit from the MMAE. This is confirmed by the scores averaged over the whole DJF season (Fig. 5b) which show no negative scores for any of the Northern Hemisphere regions, and the greatest benefits of up to 12% improvement over the EPS over Europe. In MJJ (Fig. 5c) the benefits are considerably less, with the greatest benefit observed over the N.Atlantic and Europe region and the southern hemisphere showing a small degradation in skill from EEUU54 relative to EPS at all lead-times.

(a)



(b)

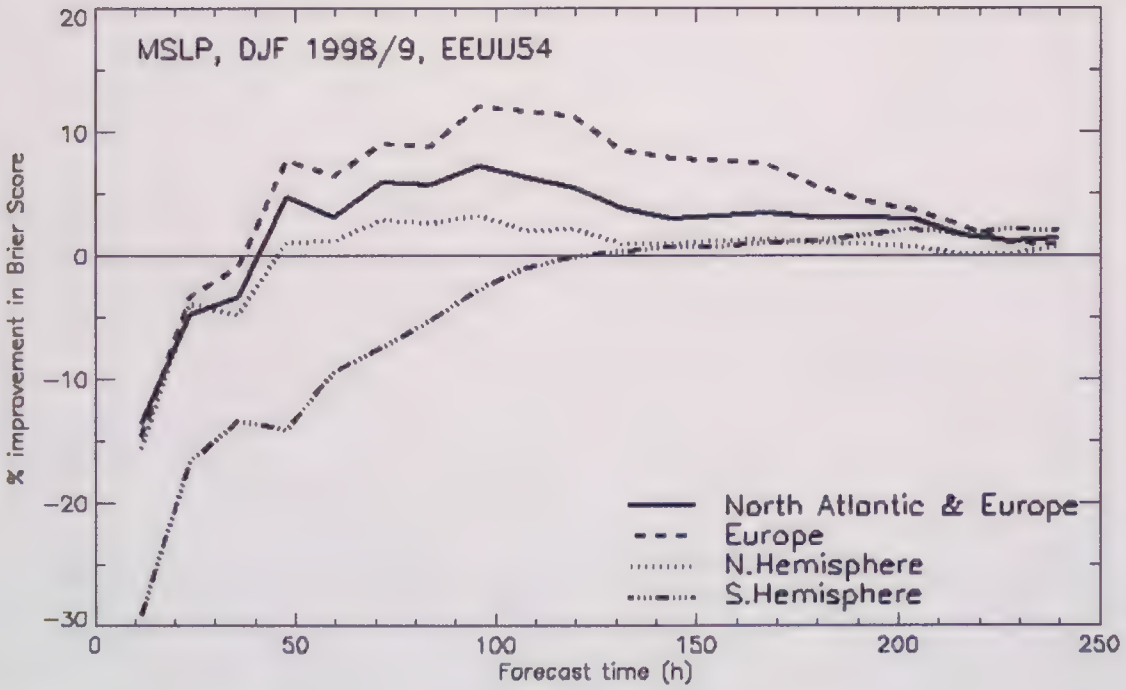


Figure 5 (contd overleaf).

(c)

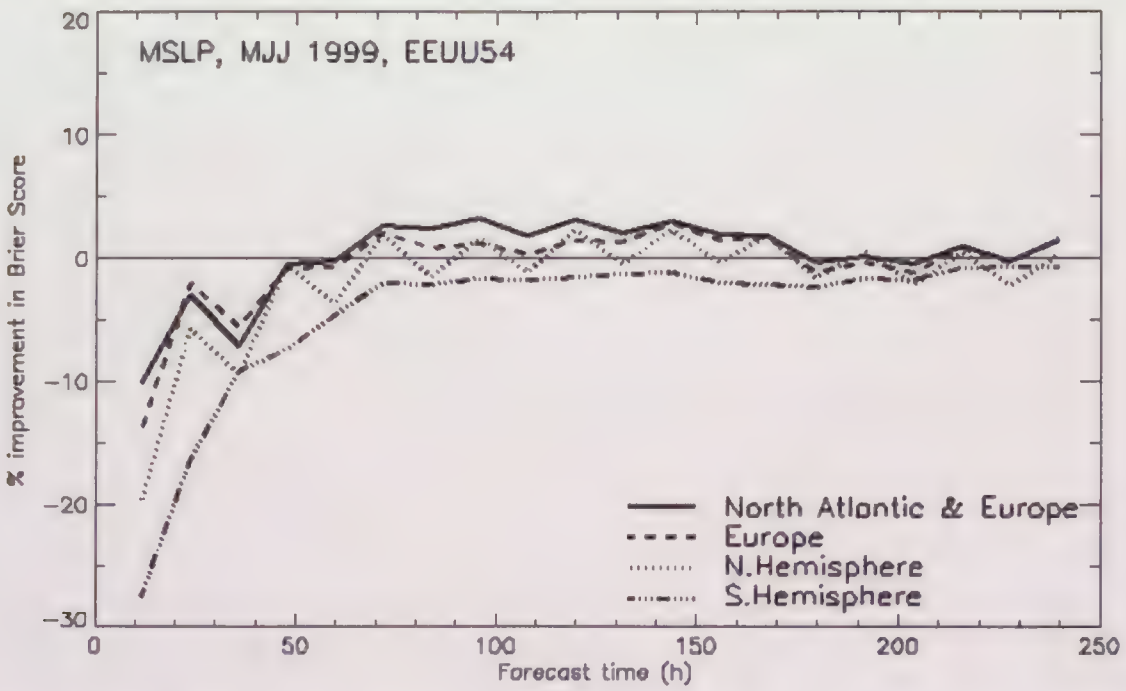


Figure 5: Brier Skill Scores relative to the EPS for MSLP from the EEUU54 ensemble, calculated over the different analysis regions as shown in the key, for (a) December 1998, (b) DJF and (c) MJJ.

It is noticeable from Fig. 5 that the MMAE benefits observed in the northern hemisphere (NH) are not achieved in the southern hemisphere (SH). It is interesting to note that in the NH the greatest gains are in the winter, whereas Fig. 5c shows no benefit from the MMAE in the southern winter (MJJ). In DJF in the SH MMAE skill is poorer than EPS through about the first 5 days of the forecast, but in the later part of the forecast there is more benefit here than in the NH. As discussed in section 4.2, and shown in Fig. 2(b), the poor performance up to T+120 is exaggerated by the use of the ECMWF analysis which has more impact in the southern hemisphere. In MJJ the MMAE is slightly less skilful than the EPS even at longer lead-times, and this is supported by verification against the Met Office analysis (not shown). There are therefore some limited benefits of the MMAE in the southern hemisphere in DJF, but not in MJJ.

These results show that the benefit of the MMAE over the EPS varies with both the time (month and day) and geographical region. This suggests that the benefit of the MMAE is provided by the different skills of the two NWP systems, and hence their ensembles, in different geographical regions and different synoptic situations. Both NWP systems used are amongst the most skilful available, but since they have entirely independent dynamical and numerical formulations it is to be expected that each will have different strengths and weaknesses. These strengths and weaknesses are likely to be flow-dependent, related to both geographical regions and synoptic types. Results show that in situations in which the ECMWF system performs less well, replacing some ensemble members with members from the UM system significantly improves the ensemble performance. This was the case in January 1999 over the North Atlantic and Europe area where the UM ensemble performed individually almost as well as the EPS, and the combination of information from both in the MMAE ensemble gave improved performance. In

situations where the ECMWF system performs well on its own, such as in December 1998 for the same region, the UM members add less benefit, but importantly they do not significantly reduce the overall skill of the ensemble. Over the more limited area of Europe the gains in December were greater, suggesting that it was largely in this area where the UM ensemble was adding useful information to the EPS. Results from the whole DJF season (Fig. 5b) show that the greatest benefits occur over Europe amongst the regions analysed. Comparing results from different months (Fig. 3) or even days (Fig. 4) shows that within a region the benefits vary considerably, and this is most probably caused by variations in the synoptic type. Considering results from the northern hemisphere, benefits from the MMAE are greater in the northern winter (DJF) than in the summer (MJJ), as seen from Figs. 1(a) and (b). In summary, most of the benefits from MMAE ensembles will be gained in regions where the EPS is performing relatively poorly at any particular time, and it can be expected that the regions where this occurs will vary on a range of time-scales from daily to seasonally.

5. DETERMINISTIC FORECAST SKILL

To obtain maximum benefit from the full information content of ensemble forecasts, they are best interpreted probabilistically, hence the use of the Brier Skill Score as the main diagnostic in this paper. However it is also common practise to use the ensemble mean (EM) as a deterministic forecast tool, and many studies (eg Molteni *et al*, 1996) have shown enhanced forecast skill from the EM compared to the control forecast. It is therefore useful also to consider the EM skill of the MMAE ensemble. A standard measure of the quality of a deterministic forecast field is the *RMSE* (root mean square error). The *RMSE* skill score is obtained from equation 1 with X being the *RMSE*.

In Fig. 6 the MSLP *RMSE* skill scores of the two configurations of the MMAE are plotted against forecast lead-time for the North Atlantic and Europe in the (a) DJF and (b) MJJ periods. Results for EM *RMSE* are similar to those for Brier Skill Scores shown in Fig. 1, although the benefits of the MMAE are slightly smaller. In DJF the MMAE ensembles both give up to 5% improvement in EM skill over the EPS for lead-times over 36 hours. In MJJ the EM skill of the MMAE is up to 3% greater than EPS between T+48 and T+144. Beyond T+144 the MMAE is no better than EPS, but it is not significantly worse. When the *RMSE* is plotted directly (not shown), rather than as a skill score, this can be seen to give a gain in predictability of the order of 6 hours between T+36 and T+144 in winter, and up to 15 hours at longer lead-times.

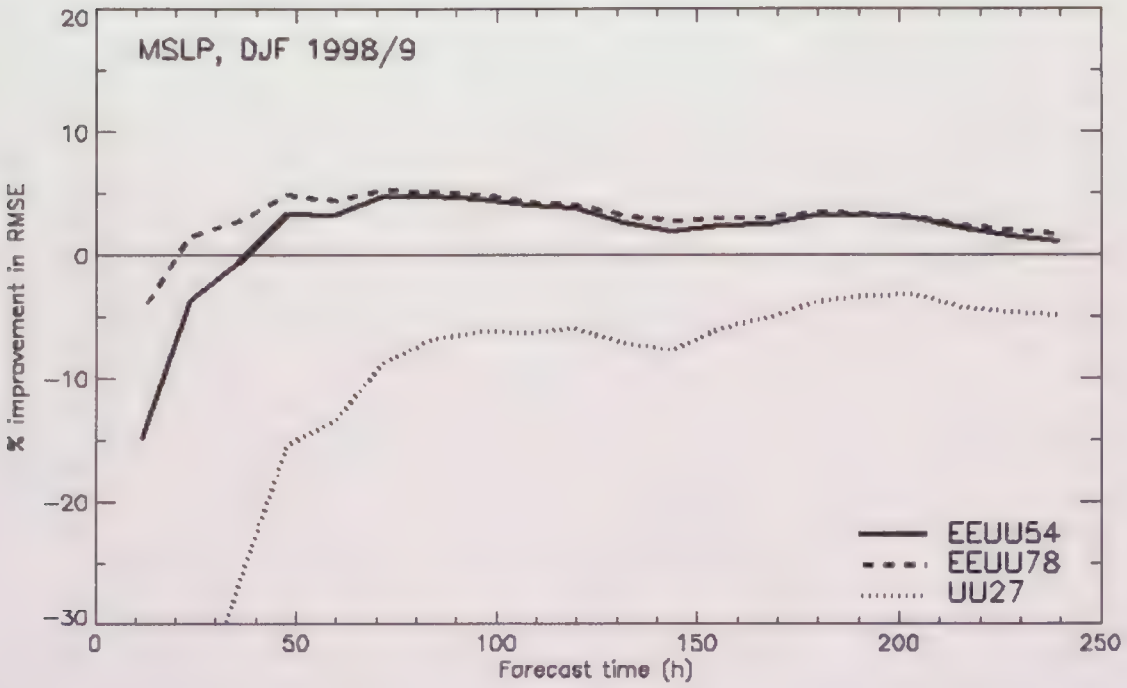
As with the Brier Score, there is little difference in results between the EEUU54 and EEUU78 ensembles, indicating that most of the gains in EM skill from the MMAE system are due to the greater effectiveness of the multi-model multi-analysis system, and not simply to increasing the ensemble size.

Results for H500 (not shown) are similar to those for MSLP, with marginally larger improvements in skill over EPS. For precipitation the gains in skill are small, up to 2%, but equally there is no degradation compared to EPS.

Variations in the benefits of the MMAE from month to month and region to region are very similar to those found in the *BSS* results. For example in January 1999 the EEUU54 ensemble was between 5 and 9% better than the EPS in EM *RMSE* for all lead-times from T+48 to T+192, but in December 1998 it was between 3% better and 1% worse for the same periods.

An alternative measure of EM deterministic skill is to use Anomaly Correlation Coefficients (*ACC*) instead of *RMSE*. Results using *ACC* (not shown here, but see Richardson 2000) are very similar to those measured with *RMSE*.

(a)



(b)

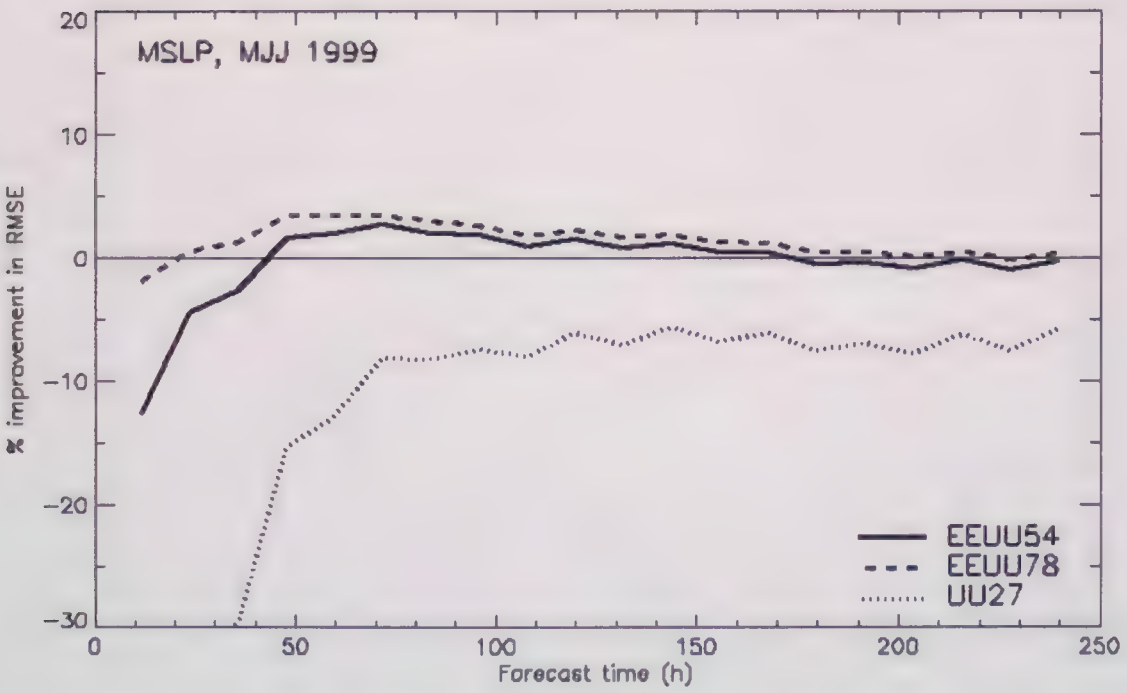


Figure 6: Ensemble Mean *RMSE* skill score of MSLP, relative to the EPS, calculated over the North Atlantic and Europe for (a) DJF and (b) MJJ. Different ensembles as in Fig. 1.

6. RELATIVE OPERATING CHARACTERISTIC (ROC)

Another measure of skill which is frequently used for probabilistic forecasts is the ROC (Relative Operating Characteristic) curve. ROC originates in signal detection theory, and was introduced into meteorology by Mason (1982). A full description of its use is given by Stanski *et al* (1989). ROC measures the skill of a forecast in terms of a hit rate (*HR*) and a false alarm rate (*FAR*) both classified according to the observations. Because it is classified by observations (as opposed to forecasts), ROC measures the ability of a forecast system to discriminate between two alternative outcomes (eg MSLP above or below normal). It therefore measures how useful the forecasts are for decision-making. Murphy (1973) decomposed the Brier Score into three terms, reliability, resolution and uncertainty. Uncertainty depends only on the observations. Reliability measures how well the forecast probabilities relate to the frequency of occurrence of the event; resolution measures how effectively the forecast system is able to discriminate occasions when probabilities are high or low. Being classified by the observations, ROC is closely related to resolution but is independent of reliability, which classifies forecasts by the forecast probability.

The ROC Hit Rate and False Alarm Rate are defined from the 2×2 contingency table for forecasts of a binary event, given in table 2, as follows:

$$HR = \frac{H}{H+M} \tag{3}$$

$$FAR = \frac{F}{F+R} \tag{4}$$

More skilful forecasts are characterised by higher hit rates and lower false alarm rates; if the system has no skill then the forecast is independent of whether the event occurs, and *HR*=*FAR*. For probability forecasts the ROC curve is generated by calculating hit rates and false alarm rates for a range of probability thresholds of an event, where the event is forecast to occur if the probability exceeds the threshold, and then plotting *HR* against *FAR*. This forms a curve from (0,0) to (1,1). For skilful forecast systems the curve is bowed towards the top left corner where *HR*=1 and *FAR*=0; for forecasts with no skill the curve lies along the line *HR*=*FAR*. The area under this curve gives a measure of the overall skill of the forecast system, with a maximum value of 1.0 for a perfect forecast and a value of 0.5 indicating no skill.

		Event Forecast	
		Yes	No
Event Observed	Yes	Hit *H*	Miss *M*
	No	False Alarm *F*	Correct Rejection *R*

Table 2: Contingency table of forecast performance. Letters *H, M, F, R* represent the total numbers of occurrences of each contingency in the verification sample.

Skill scores relative to EPS may be calculated for area under the ROC curve using equation (1). ROC area skill scores of the UU27, EEUU54 and EEUU78 ensembles for MSLP above normal over the North Atlantic and Europe are plotted for January 1999 in Fig. 7. Similar scores for EEUU54 only are plotted in Fig. 8 for individual months of (a) DJF and (b) MJJ. Mean values for the DJF and MJJ seasons are also plotted in Fig. 8, although it should be noted that these are simple arithmetic means of the ROC skill scores for the three individual months. (It was not possible to calculate seasonal mean skill scores from ROC *HR*s and *FAR*s for the full three-month seasons due to the way data was stored.) Results for EEUU54 may be compared with *BSS* results given in Figs. 1 and 3, and there are some important differences. As with the Brier Scores, the greatest benefits of the MMAE are in the first 5 days of the forecast. Percentage skill improvements over the EPS in this period were greater in terms of ROC area than in Brier Scores: in January gains of 10-22% were obtained, and in December and February 5-14%. Mean benefits in DJF were 10-15% compared to 3-7% as measured by *BSS*, and in MJJ 5-7% compared to 2-3% for *BSS*. Beyond T+120 the benefit decreased, particularly in the winter DJF season, but apart from beyond T+192 in DJF, the seasonal mean skill scores were all positive. Negative skill scores occurred for some lead-times in some months, most notably in December with negative scores of 5-10%. There was little difference between EEUU54 and EEUU78 (Fig. 7), although the difference was slightly greater in some other months such that skill scores for EEUU78 in December remained positive out to T+180 and then went no lower than -4%.

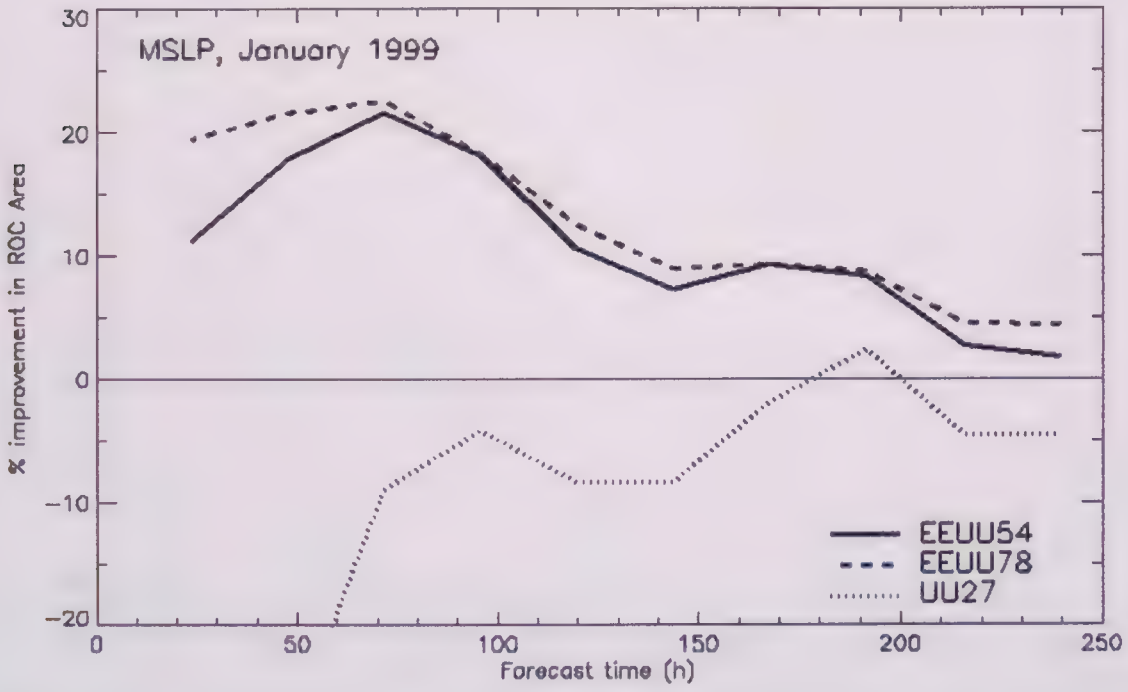
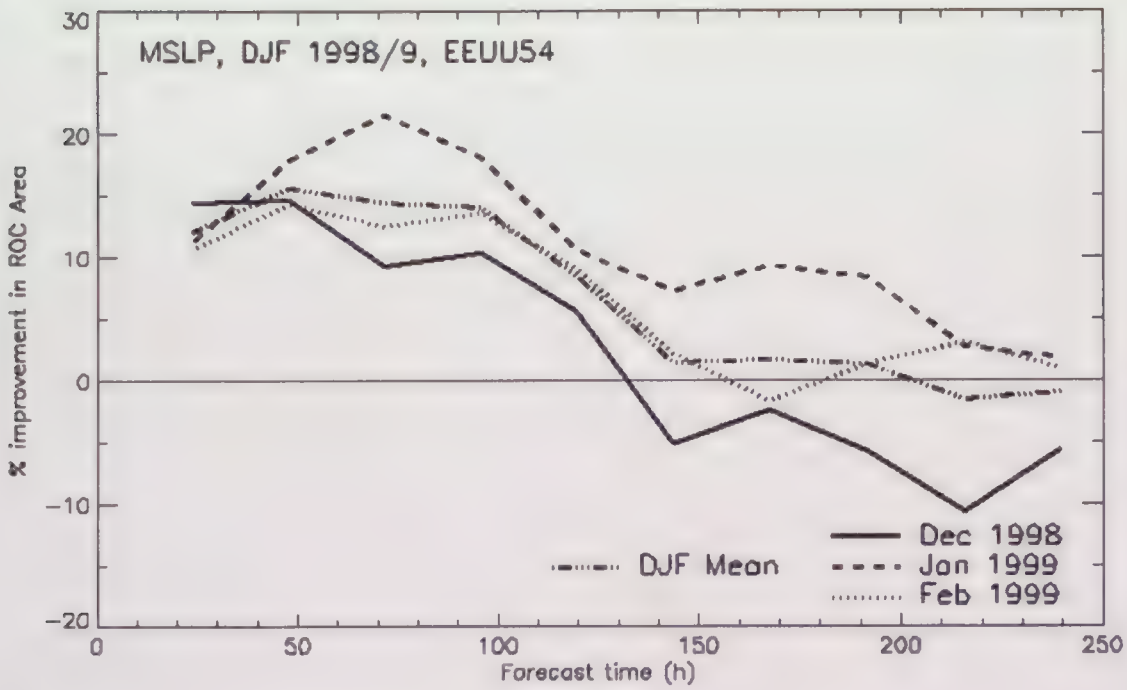


Figure 7: Area under ROC skill scores, relative to EPS, for MSLP above normal over the North Atlantic and Europe in January 1999 for different ensemble configurations as given in the key.

(a)



(b)

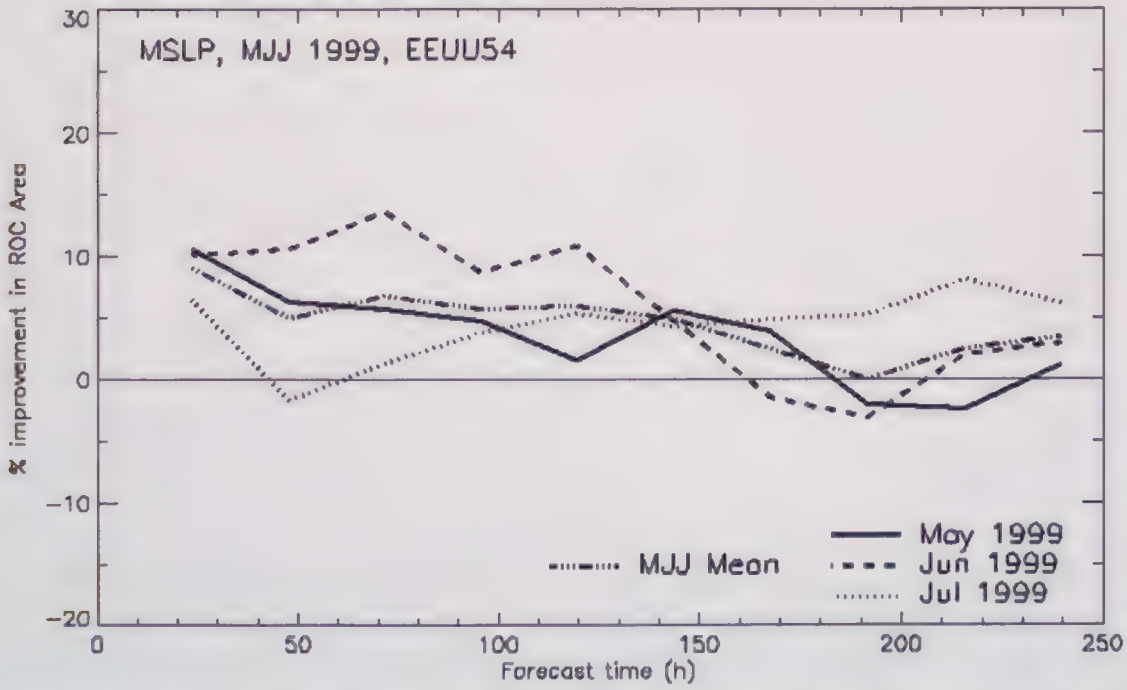


Figure 8: Area under ROC skill scores, relative to EPS, for MSLP above normal from the EEUU54 ensemble over the North Atlantic and Europe in (a) DJF and (b) MJJ. Scores are shown for individual months and seasonal means as shown in the key. (Note: the seasonal mean values are simple arithmetic means of the skill scores for the three months.)

7. ENSEMBLE SPREAD

It was noted in the introduction that the spread of the EPS is not sufficient to cover the full range of forecast uncertainty. Evans *et al* (2000) demonstrated that for their case studies the MMAE substantially reduced the proportion of observations lying outside the spread of the ensemble, indicating an improved coverage of the range of uncertainty. The spread of each ensemble configuration is illustrated in Fig. 9 for DJF over the N.Atlantic and Europe. Spread is here measured as the RMS spread about the ensemble mean. Spread of the MMAE configurations is increased relative to the EPS spread at all lead times. Initial spread is enhanced due to the inclusion of different analyses. The EPS initial perturbations are purposely small since the singular vector (SV) approach produces perturbations which grow very rapidly over the early part of the forecast. The differences between ECMWF and UM analyses may better reflect the true uncertainty in the initial conditions, although the difference will not grow nearly so rapidly as the SV perturbations. The enhanced spread persists throughout the forecast and grows slowly in absolute terms (hPa) although it decreases as a proportion of the total spread. There is little difference in spread between the EEUU54 and EEUU78 ensembles, showing that the spread is enhanced by the inclusion of ensemble members from a different NWP system, and not simply by increasing the number of members. The spread of the UM ensemble (UU27) grows slightly more slowly than that of the EPS, and this reflects the fact that the SV perturbations are calculated to produce maximum ensemble growth with the ECMWF model, and also that the EPS includes stochastic physics perturbations which are not present in the UM ensemble. Very similar results are obtained in the MJJ season, except that overall growth rates are slower in the summer.

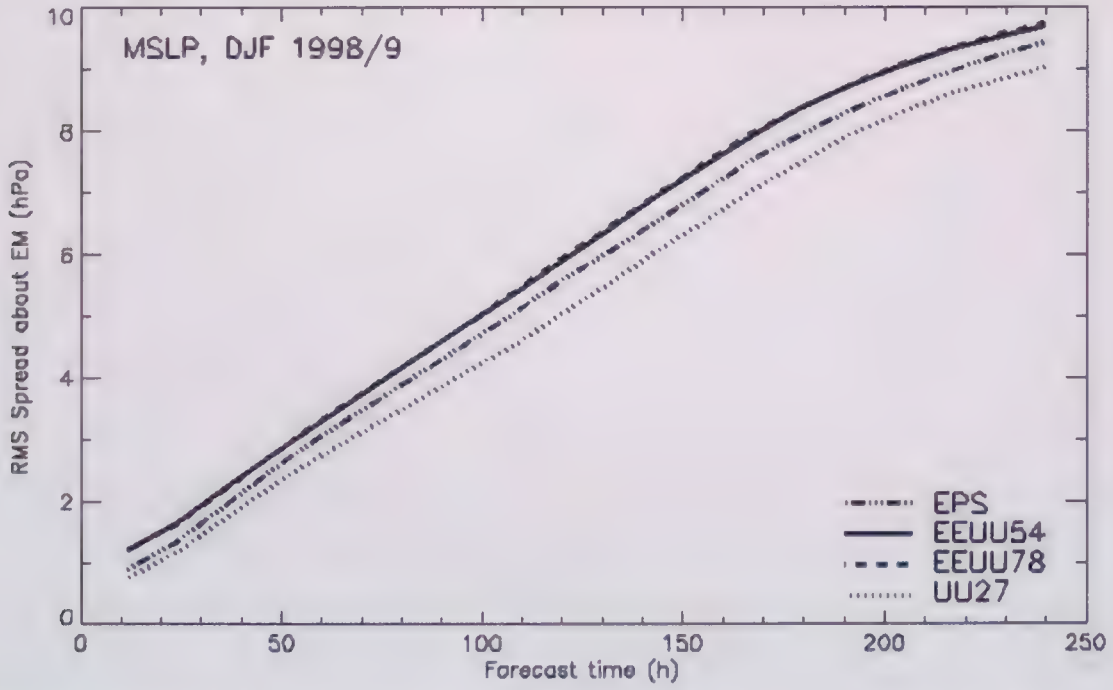


Figure 9: RMS spread of the ensemble around the ensemble mean, for MSLP over the North Atlantic and Europe in DJF for different ensemble configurations as given in the key.

8. DISCUSSION AND CONCLUSIONS

Results presented above have illustrated that considerable gains in ensemble forecasting skill can be gained by use of an MMAE approach. With no increase in overall ensemble size, improvements of 3-7% in Brier Skill were achieved in DJF, with 10-15% improvements in area under ROC. Benefits were rather less in the early summer season MJJ, but nevertheless mostly positive. Occasions when the MMAE performed less well than the EPS were unusual, and restricted to a small proportion of lead-times - in almost all cases analysed for MSLP, H500 and P24 the benefits are far greater than any reductions in skill. The only real exception to this was for MJJ in the southern hemisphere where the EEUU54 ensemble performed slightly worse than the EPS at all lead-times.

In the introduction it was noted that Buizza *et al* (1999) have introduced stochastic physics into the EPS in order to represent some aspects of model uncertainties. Benefits of this change have mostly been limited to improvements in verification of weather parameters, particularly precipitation. This is perhaps to be expected, as the feedback from the physics to the dynamics of the model is relatively weak. The stochastic physics, along with the introduction of evolved singular vectors as part of the initial condition perturbations (Barkmeijer *et al*, 1999; Buizza *et al*, 2000) has also resulted in an increase of about 4% in ensemble spread. These changes were introduced to the EPS before the current study was conducted, so the benefits of the multi-model ensembles demonstrated here have been achieved over the EPS with the stochastic physics system in operation. This demonstrates that, although the stochastic physics scheme is beneficial, there are still substantial further benefits to be gained by addition of one or more other models with different, skilful attractors, along with their analyses, into the ensemble system.

As discussed in section 4.3, the benefits of the MMAE are flow-dependent, and vary both in time and geographically. In comparison with a single-model ensemble (such as the EPS), it could be expected that the MMAE might perform better in synoptic situations which are generally handled better by the additional model. However the differences in performances of skilful models are often subtle, and the situations in which one or the other is likely to perform better are normally impossible to identify synoptically. By combining the two NWP systems together into a joint ensemble, the results demonstrate that the benefits of whichever is the better system at a particular time and place may be obtained all the time. On both monthly and daily time-scales, the MMAE ensembles thus effectively act as a filter for the better-performing individual ensemble, and on occasions perform better than either individual ensemble. An important result of this study is that even when the single model ensemble is performing well, and the second model less well, as in the December 1998 results, the joint ensemble does not significantly degrade the performance of the single-model ensemble. Again, the only real exception to this was for MJJ in the southern hemisphere.

In this study considerably greater benefits of the MMAE were observed for the winter DJF period than for the early summer MJJ period. This is probably due mostly to the different behaviour of the atmosphere in summer and winter, and is likely to be a general result typical of any year. However it is interesting to note that separate studies of the EPS have shown that it performed exceptionally well in the summer of 1999 compared to other recent years (Roberto Buizza, personal communication). This good EPS performance would have made it a particularly difficult period for the MMAE to improve upon, particularly given that we find that the MMAE often picks out the performance of whichever is the best individual system at any one time. It might therefore be expected that in more typical summers there would be a greater advantage to be gained by the use of an MMAE system. This may depend on whether the unusually good performance was related to the ECMWF model or to the perturbations. Since the perturbations were applied to both the EPS and UM ensembles, both components would benefit from unusually

good perturbations and the results from this study are likely to be typical; if it was the ECMWF model performing exceptionally well then it is likely that in other years the MMAE would give greater benefit. In fact the EPS control performed poorly in summer 1999, so the good EPS performance was probably due to the perturbations, and the MMAE results are likely to be reasonably typical.

In terms of the geographical regions examined, the greatest benefits of the MMAE were over Europe in the winter DJF period, and over the N.Atlantic/Europe region in the summer MJJ period. If gains are due to the models' different performances for synoptic-scale systems, it might be expected that on occasions larger benefits will be observed over small geographical regions when particular synoptic types prevail in those regions. Results observed are consistent with this idea.

As noted earlier, ROC is related to the resolution of the forecasts, and measures the ability of a forecast system to aid decision-making. The fact that in the first 5 days of the forecast ROC area skill scores for the MMAE are greater than the Brier Skill scores suggests that the improvements in probabilistic forecast skill are mostly in improved resolution of the forecasts. This would be consistent with the idea that incorporating extra members into the ensemble from a different NWP system would increase the chance of the ensemble including solutions which are more synoptically different from each other. As noted in the introduction, Met Office forecasters often observe that the EPS does not spread sufficiently to incorporate the full range of uncertainty, and that model forecasts from other NWP centres are more different synoptically from the ECMWF model than are the EPS members. The use of an MMAE ensemble brings some of the extra information available from other NWP systems into the ensemble, and results in a greater spread in the forecast solutions, as seen in Fig. 9. This shows that the addition of members from another NWP system is an effective way to introduce new directions of growth into the ensemble. These new directions originate from the different attractor of the additional model, and manifest themselves through both forecast and analysis differences. Of course for the additional members to add useful information, the NWP system from which they are derived must be of comparable skill to the original system in the ensemble, although as demonstrated here it need not be quite as skilful provided that it is better in some locations on some occasions.

Although the amounts of data analysed were large, the numbers of forecasts used in this study are still relatively small (75 days in DJF and 85 days in MJJ). Nevertheless they are very much greater than has been possible in previous studies such as Evans *et al* (2000) which used nine case studies. Variations in MMAE benefits from month-to-month make it difficult to draw overall conclusions from monthly results. However results for the full DJF and MJJ seasons (Figs. 1, 5b, 5c, 6 and 8) generally show much more consistent results at different forecast lead-times than are seen for individual months, and this gives some confidence that seasonally averaged results are reasonably statistically stable, and significant. The MMAE benefits found here are mostly rather smaller than those reported by Evans *et al* (2000). For example Evans *et al* (2000) reported Brier Skill improvements over the North Atlantic and Europe of around 9% for H500, whereas results reported here are 3-7% improvements. This is to be expected since Evans *et al* (2000) used case studies, many of which were chosen because of poor EPS performance, whereas this paper used all available forecasts through the seasons analysed. The fact that the MMAE continues to give benefits in this quasi-operational environment strongly reinforces the results of Evans *et al* (2000) and demonstrates that benefits are not restricted to occasional days of poor EPS forecasts.

In section 2 it was noted that the use of the EPS perturbations to create the UM ensemble was likely to produce a less optimal performance from the UM ensemble than from the EPS, since the perturbations are specifically calculated to generate maximum ensemble growth using the

ECMWF model. This is borne out by the results which show (Figs. 1, 6 and 7) that the UM ensemble alone was very poor compared to the EPS by every measure used. This poor performance was not simply due to the smaller ensemble size, since a reduced EPS consisting of the 27 members using the same perturbations as the UM ensemble (not shown) performed only marginally less well than the full EPS. As well as the fact that the initial perturbations are optimised for the ECMWF model and analysis, the poorer performance of the UM ensemble is also partly due to the lack of stochastic physics perturbations which are included in the EPS. Despite this poor performance of the UM component, the MMAE still provides considerable benefits. It may therefore be speculated that even more benefits might be gained from the combined ensemble if the UM component was better optimised, by including both specially generated perturbations and a stochastic physics system. In addition further gains in ensemble skill could be achievable by adding more components to the MMAE, with members derived from other independent NWP systems.

Analysis of probabilistic skill in this paper has been limited to common events such as 'MSLP above normal'. It has not been possible to include more extreme events such as 'MSLP more than one standard deviation above/below normal'. It is generally found that more extreme events are more difficult to forecast, either deterministically or probabilistically. However there is much interest in whether ensembles can be used for forecasting the probabilities of extreme events, particularly for forecasting severe weather. Richardson (2000) included some analysis of the MMAE skills for more extreme events for H500 (25, 50 and 100m above and below normal) and T850 (4 and 8K above and below normal). In terms of *BSS* he found similar benefits from the MMAE for these more extreme events, at least indicating that the MMAE improves the probabilistic prediction of more extreme events as well as less extreme ones. Using area under ROC the skill scores were generally higher for the more extreme events. For example for 3-10 day forecasts of H500 over Europe Richardson (2000) reported a ROC skill score of 6.03 for 'above normal', but for '+100m above normal' ROC skill was 7.43 and for '-100m below normal' it was 10.52. This indicates that the MMAE improves the ability of the ensemble to resolve when extreme events are more or less likely to occur. Since ROC measures how useful the forecasts are for decision-making, this suggests that the MMAE may be particularly valuable for forecasting more extreme (severe) weather.

One aspect of the MMAE system which has not been addressed above is how much of the benefits of the MMAE are due to incorporating the additional analysis and how much to the additional model. Evans *et al* (2000) conducted a more detailed study of the relative importance of analysis and model dependencies by including ensembles run using the UM initialised with the ECMWF analysis. They concluded that the effects of including more than one model were non-negligible in the first 48 hours, become equal with analysis dependencies around Day 8 of the forecast, and dominate thereafter. From this it is likely that a significant part of the benefits of the MMAE described above are due to the additional analysis rather than the additional model. Richardson (2000) has presented a comparison of results of the MMAE with a multi-analysis ensemble generated using the ECMWF model, and concluded that the multi-analysis ensemble gave 50-80% of the benefits of the MMAE as measured by BSS and 70-80% in ROC terms. Richardson goes on to demonstrate that if model biases are first removed before results are analysed, the multi-analysis ensemble becomes more competitive with the MMAE. Nevertheless the MMAE remains the most skilful system overall, and in any case bias correction is not always a viable option in operational systems where biases change as model upgrades are introduced on a relatively frequent basis. Thus, while substantial benefits may be gained by incorporating multi-analyses, the multi-model approach is still required for the full benefits of the MMAE system.

Previous work has demonstrated the benefits of multi-model multi-analysis ensembles for limited case studies (eg Evans *et al*, 2000). The aim of this study was to determine the extent to which such benefits may be obtained in a quasi-operational system using the current operational versions of models. Analysis of the benefits of the MMAE approach has primarily used Brier Skill Scores as an overall assessment of the probabilistic ensemble skill. Ensemble mean *RMSE* and ROC have also been used with broadly similar conclusions. Results shown have illustrated that substantial gains in forecast skill may be achieved by adding a second skilful model and analysis into the EPS system, and that in most cases these gains may be achieved without increasing the overall size (and therefore running cost) of the ensemble. Benefits were greatest in the northern winter (DJF) over Europe, but lesser benefits were gained over most regions in most seasons. Gains achieved are somewhat less than observed by Evans *et al* (2000) using case studies, but are nevertheless consistently achieved. Evidence from variations in the benefit of the MMAE suggests that gains in skill come from the additional model performing better in certain synoptic situations. In constructing an MMAE system for optimal performance it will therefore be important to ensure that as well as being of similar overall skill, the NWP systems combined are also as independent of each other as possible so that their errors and weaknesses are as independent as possible. This will maximise the chance that when one element of the system performs poorly, another element will do better

Multi-model multi-analysis ensembles are now becoming the standard technique for seasonal range forecasting (Graham *et al*, 2000), and are also being used increasingly for short-range forecasting (Stensrud *et al*, 1999; Mullen *et al*, 1999). Results presented here, along with those of Evans *et al* (2000) and Richardson (2000), demonstrate that substantial improvements can also be gained from the MMAE approach in medium-range forecasting.

ACKNOWLEDGEMENTS

The authors would like to thank Mike Harrison, currently of WMO in Geneva, for his support in initiating this project. We would also like to thank Richard Barnes for his immense efforts in installing the UM on the ECMWF computer system, and setting up and running the UM ensemble system. Useful support was also provided by Kelvyn Robertson and Anette van der Wal. Many useful discussions have been held with Tim Palmer and David Richardson of ECMWF.

REFERENCES

Barkmeijer, J., Buizza,R., and Palmer,T.N., 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System *Q. J. R. Meteorol.. Soc*, **125**, 2333-2351.

Brier,G.W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* **78**, 1-3.

Buizza,R., Barkmeijer,J., Palmer,T.N. and Richardson,D., 2000: Current status and future developments of the ECMWF Ensemble Prediction System. *Meteorol. Appl.* **7**, 163-175.

Buizza,R., Miller,M. and Palmer,T.N., 1999: Stochastic representation of model uncertainties in the ECMWF EPS *Q. J. R. Meteorol.. Soc*, **125**, 2887-2908.

Buizza,R. and Palmer,T.N., 1995: The singular-vector structure of the atmospheric general circulation. *J. Atmos. Sci.* **52**, 1434-1456.

Buizza,R., Petroliagis, T., Palmer,T.N., Barkmeijer,J. Hamrud, M., Hollingsworth, A., Simmons, A., and Wedi, N. , 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system *Q. J. R. Meteorol.. Soc*, **124**, 1935-1960.

Cullen,M.J.P., 1993: The unified forecast/climate model. *Meteorol. Mag.* **122**, 81-93.

Evans,R.E., Harrison,M.S.J. and Graham,R.J., 2000: Joint Medium Range Ensembles from the UKMO and ECMWF Systems. *Mon. Wea. Rev.* **128**, 3104-3127.

Houtekamer,P.L., Lefaivre,L., Derome,J., Ritchie,H., and Mitchell,H.L., 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**, 1225-1242.

Graham,R.J., Evans,A.D.L., Mylne,K.R., Harrison,M.S.J. and Robertson,K.B., 2000: An assessment of seasonal predictability using Atmospheric general circulation models *Q. J. R. Meteorol.. Soc*, **126**, 2211-2240.

Molteni,F., Buizza,R., Palmer,T.N. and Petroliagis,T., 1996: The ECMWF Ensemble Prediction System: Methodology and Validation. *Q. J. R. Meteorol.. Soc*, **122**, 73-119.

Mason,I., 1982: A model for the assessment of weather forecasts. *Aust. Meteorol. Mag.* **30**, 291-303.

Mullen,S.L., Du,J. and Sanders,F., 1999: The dependence of ensemble dispersion on analysis-forecast systems: implications to short-range ensemble forecasting of precipitation. *Mon. Wea. Rev* **127**, 1674-1686.

Murphy,A.H., 1973: A new vector partition of the probability score *J.Appl.Meteorol.*, **12**, 595-600.

Richardson,D, 2000: Ensembles using multiple model and analyses, submitted to *Q. J. R. Meteorol.. Soc*, January 2000. (Also in: Proceedings of a seminar held at ECMWF on diagnosis of models and data assimilation systems, 6-10 September, 1999, Shinfield Park, Reading.)

Stanski,H.R., Wilson,L.J. and Burrows,W.R., 1989: Survey of Common Verification Methods in Meteorology, WMO WWW Tech. Report No 8, WMO TD No 358.

Stensrud,D.J., Brooks,H.E., Du,J., Tracton,M.S. and Rogers,E., 1999: Using Ensembles for Short-Range Forecasting *Mon. Wea. Rev.*, **127**, 433- 446.

Toth,Z., and Kalnay,E., 1993: Ensemble Forecasting at the NMC: The generation of perturbations. *Bull. Amer. Meteorol. Soc.*, **74**, 2317-2330.

Wilks,D.S., 1995: Statistical Methods in the Atmospheric Sciences - An Introduction, International Geophysics Series Vol 59, Academic Press.